SUPPLEMENTAL FILE 1. STATISTICAL ANALYSIS IN GREATER DETAIL AND JUSTIFICATION OF *k*-VALUE

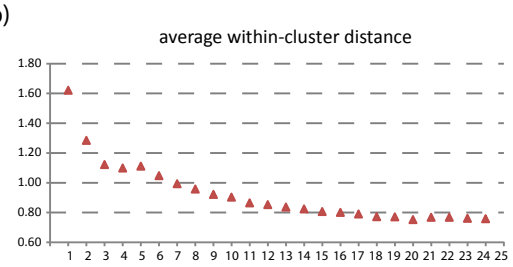
In order to investigate the effect of the selected number of clusters, *k*, on the statistical results and to evaluate the impact on the ability to distinguish the 10 major lithological classifications of the New Jersey successions, a variety of tables and plots are presented in Supplemental File 1. Table (a) compares average within-cluster distances and between-cluster distances (summed deviation about seed points) for $k = 2$ to $k = 25$. Due to the gradational nature of the spectral gamma-ray data set presented in this paper, the selected choice of *k* is not optimal in the true sense of the word and, although formal methods of determining *k* exist, there is debate over the most appropriate method (e.g., Pham et al., 2005). Less formal methods involve either a simple visual choice of the most appropriate *k* or an analysis of changes in the within-cluster distances or other empirical measures with increasing *k*: see columns with gray text and shown graphically in plots (b) to (d). Average within-cluster distances decrease with increasing *k* and the summed deviation about seed points multiplied by *k* increases. In both cases, break-points could indicate the optimum number of clusters, whereas where values become nearly constant this suggests that the optimum number of clusters has been reached prior to this point. Plot (d) shows the derivative of the difference between successive summed deviations with the larger differences maybe indicating the optimum number of clusters. From these plots, a likely possibility for the optimum number of clusters is at $k = 4$. However, due to the ambiguity of this in such data sets and also the hypothesis in this paper that the 10 major lithological classifications can be identified from a statistical analysis, the remaining data in Supplemental File 1 compare the $k = 10$ cluster example to the $k = 4$ cluster example.

Tables (e) and (g) and their corresponding three-dimensional plots (f) and (h) compare cluster characteristics where $k = 10$ and $k = 4$. Table (i) lists four key characteristics in terms of K, U, and Th concentrations with the clusters that correspond to these alongside for both examples. Note that the comparisons between clusters of the two examples should be considered a guide only (gray text) rather than a formal statistical assessment. Plot (j) is a comparison of lithology and clusters downhole for $k = 10$ and $k = 4$ (colors as in the three-dimensional plots). These direct comparisons between $k = 10$ (used herein) and $k = 4$ highlight the benefits of using $k = 10$ for the lithological comparisons presented here (see key to starred intervals), particularly to diagnose those sediments containing glauconite.

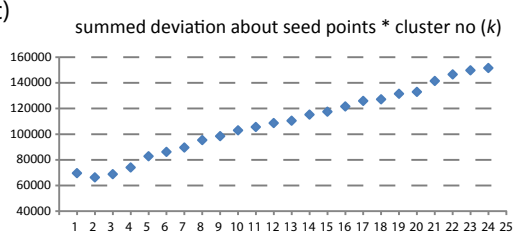
(a)

number of clusters, *k*	average within -cluster distance	summed deviation about seed	summed deviation * k	deviation for (k=n+1) - (k=n)	value of derivative (k=n)/ (k=n+1)
1	na	na	na	na	na
2	1.62	34800	69600	-12700	2.59
3	1.28	22100	66300	-4900	2.04
4	1.12	17200	68800	-2400	2.40
5	1.10	14800	74000	-1000	0.67
6	1.11	13800	82800	-1500	1.36
7	1.05	12300	86100	-1100	1.83
8	0.99	11200	89600	-600	0.79
9	0.96	10600	95400	-760	1.58
10	0.92	9840	98400	-480	0.86
11	0.90	9360	102960	-560	1.27
12	0.86	8800	105600	-440	0.94
13	0.85	8360	108680	-470	2.24
14	0.84	7890	110460	-210	0.62
15	0.82	7680	115200	-340	1.79
16	0.81	7340	117440	-190	1.19
17	0.80	7150	121550	-160	0.53
18	0.79	6990	125820	-300	2.50
19	0.77	6690	127110	-120	0.50
20	0.77	6570	131400	-240	-2.40
21	0.75	6330	132930	100	-1.67
22	0.77	6430	141460	-60	0.46
23	0.77	6370	146510	-130	0.72
24	0.76	6240	149760	-180	0.03
25	0.76	6060	151500	-6060	na
	graph (b)		graph (c)		graph (d)

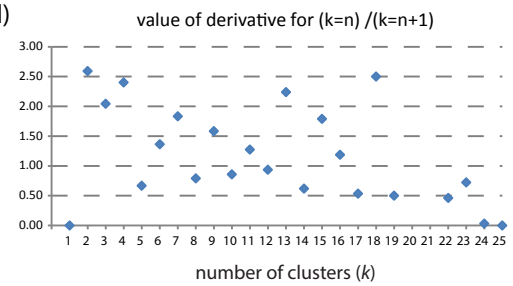
(b)



(c)



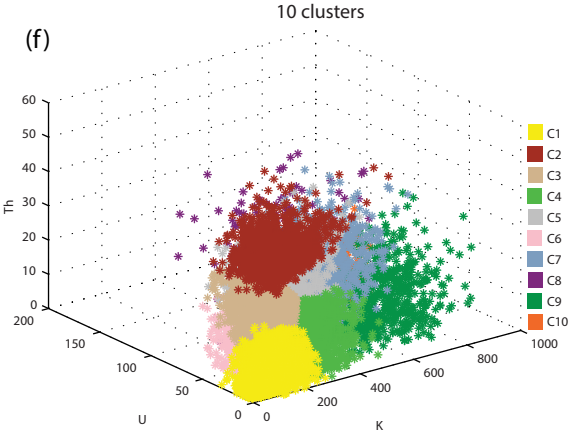
(d)



(e)

10 cluster example (*k* =10)	Number of data in cluster	Within-cluster mean distance	Cluster centroids K (Bq/kg)	U (Bq/kg)	Th (Bq/kg)
C1	4697	0.59	87	8	6
C2	1218	0.86	161	19	38
C3	2930	0.64	149	26	21
C4	1690	0.79	273	20	12
C5	915	0.92	268	43	24
C6	1783	0.75	169	38	12
C7	716	1.16	501	64	21
C8	157	1.60	520	110	22
C9	601	1.21	584	30	14
C10	2084	0.70	256	17	27
Mean	*1679*	*0.92*	*201*	*24*	*17*
		Delta	*3.11*	*3.69*	*2.2*

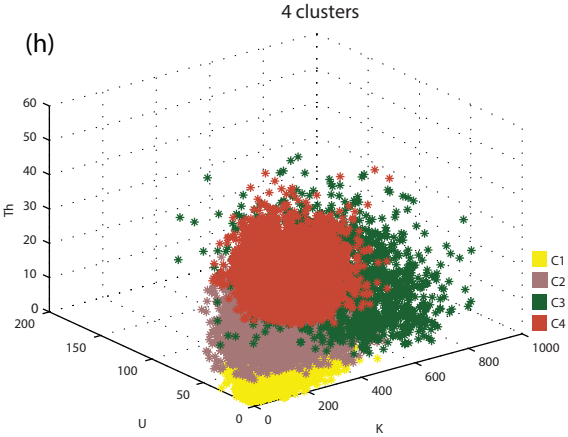
(f)



(g)

4 cluster example (*k* = 4)	Number of data in cluster	Within-cluster mean distance	Cluster centroids K (Bq/kg)	U (Bq/kg)	Th (Bq/kg)
C1	5272	0.69	99	9	6
C2	5735	0.94	204	31	16
C3	1588	1.87	520	56	19
C4	4196	0.98	204	20	30
Mean	*4198*	*1.12*	*201*	*24*	*17*
		Delta	*2.53*	*2.13*	*2.2*

(h)



(i)

cluster characteristics	corresponding cluster in *k* = 4 example	corresponding cluster(s) in *k* = 10 example
K, U and Th all low	C1	C1
K, U and Th moderate	C2	C3,C4,C5,C6
K and/or U high	C3	C7,C8,C9
Th high	C4	C2,C10

(j)